**REPORT:** STOCK PRICE HPG PREDICTION FROM HOA PHAT GROUP IN THE MANUFACTURING SECTOR AND LISTED ON THE HO CHI MINH STOCK EXCHANGE (HOSE)

**GROUP:** MLTada

**1.Exploratory Data Analysis**

**1.1. Descriptive Analysis**

**Table 1: Descriptive Statistics**

|  | **Date** | **Price** | **Open** | **High** | **Low** | **Vol.** | **Change %** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 1308 | 1308.000000 | 1308.000000 | 1308.000000 | 1308.000000 | 1,31E+09 | 1308.000000 |
| **mean** | 2020-08-18 12:58:20 | 19249.098318 | 19261.464985 | 19541.894801 | 18986.827905 | 2,43E+13 | 64.274 |
| **min** | 2018-01-02 00:00:00 | 7411.800000 | 7000.000000 | 7640.600000 | 6977.200000 | 3,32E+12 | -7.000000 |
| **25%** | 2019-05-05 06:00:00 | 11605.100000 | 11613.900000 | 11736.650000 | 11445.475000 | 1,32E+13 | -1.132500 |
| **50%** | 2020-08-18 12:00:00 | 14957.300000 | 14944.750000 | 15183.600000 | 14719.000000 | 2,10E+13 | 0 |
| **75%** | 2021-12-06 06:00:00 | 24003.825000 | 24081.000000 | 24339.325000 | 23621.425000 | 3,12E+13 | 1.222500 |
| **max** | 2023-03-31 00:00:00 | 43895.800000 | 43895.800000 | 44198.500000 | 43517.400000 | 1,10E+14 | 6.940000 |
| **std** | NaN | 9780.769912 | 9809.760926 | 9929.208070 | 9676.520954 | 1,50E+13 | 2.374579 |

**Source**: The author, 2024

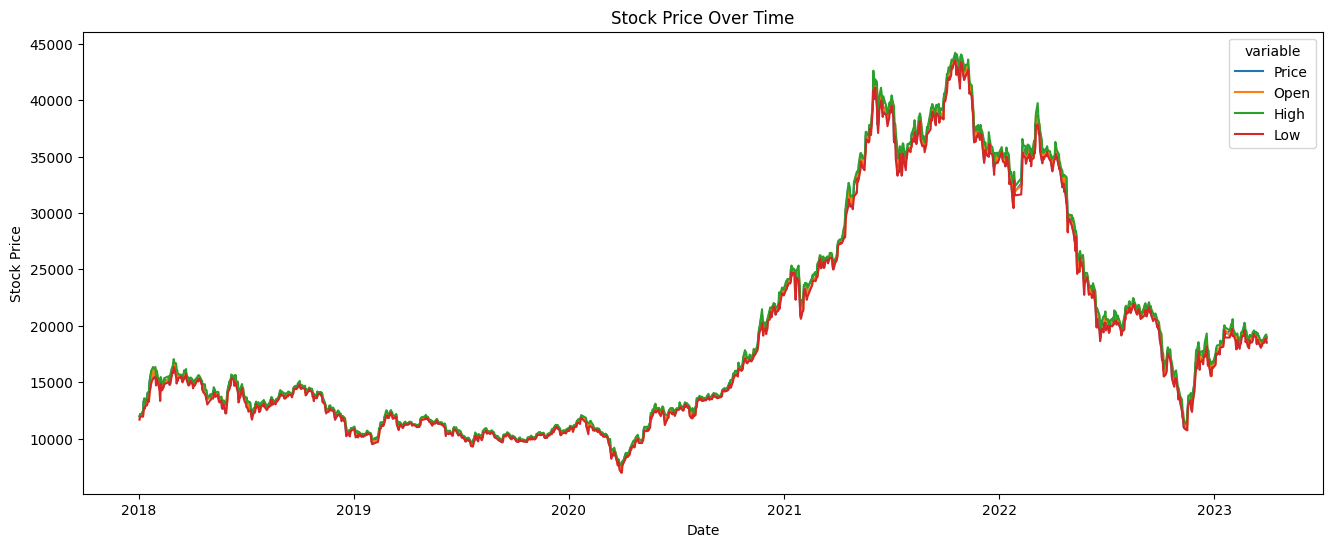
Examining the dataset, the authors gain insights into the "Price" variable. On average, the stock price of HPG is approximately 19249 VND, with a significant variation indicated by a standard deviation of 9780 VND.

To better understand the price distribution, the authors look at the quartiles. The median price figure is 19249 VND, showing that half of the stock generates below this amount while the other half exceeds it. Notably, there are outliers, with close prices ranging from a minimum of 7411 VND to a maximum of 43895 VND. These extremes may result from specific company circumstances, highlighting the diverse issues that businesses face.

**1.2. Target variable**

In this section, our team addresses the variable that the analysis team focuses on during the data analysis process, which is stock price prediction.

**Chart 1.1: Stock Price Over Time**

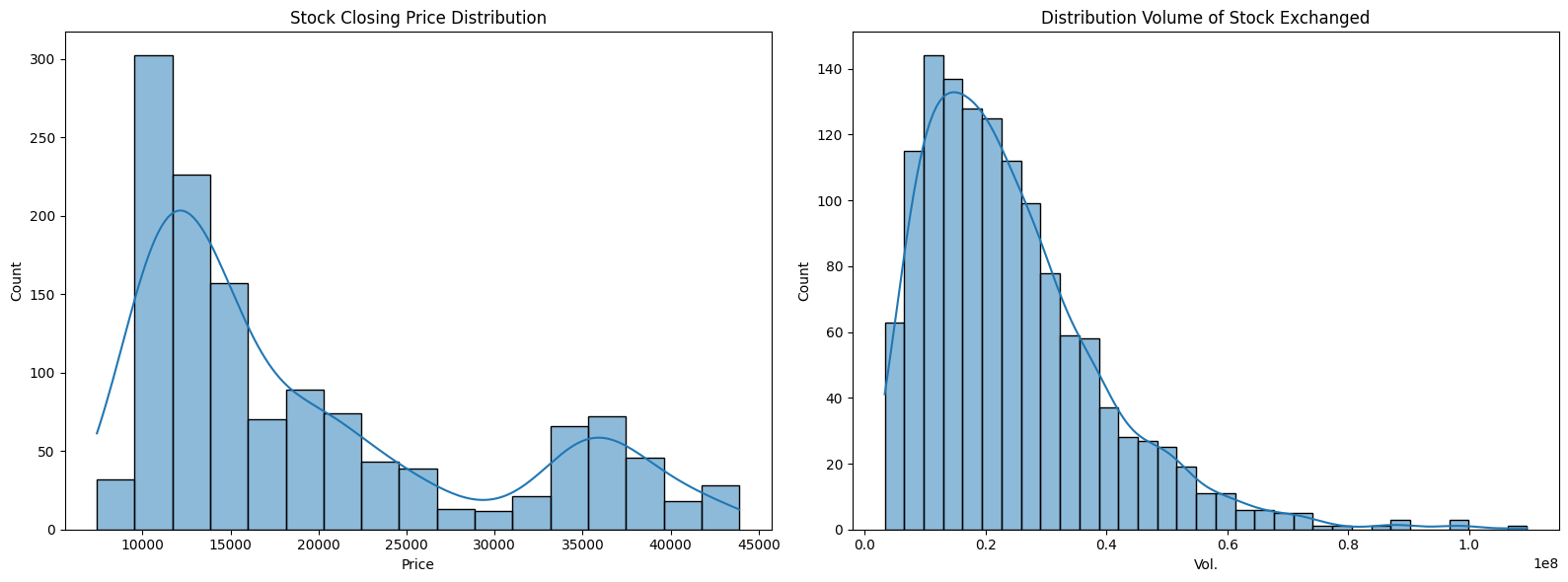
****

**Source**: The author, 2024

Based on the chart above, we can see that revenue fluctuated a lot from 2018 to 2023, and there must be many factors affecting this fluctuation. Let's analyze this fluctuation more closely through the analysis of the reasons below and build a model to evaluate and predict the close price of HPG stock.

**1.3. Univariate Analysis**

In this section, our group will show the distribution chart of the two continuous variables : **Closing Price and Volume.**

**Chart 1.2. Distribution of Closing Price, Volume of Stock Exchanged**

**Source**: The author, 2024

Based on the charts above, we can see that the distribution is skewed to the left, so the majority of stock price per day is low, around 10000 to 25000 VND; however, there are still many days where stock price is at 25000 VND or higher or more, especially many days with stock price of more than 40000 VND. Therefore, the price distribution is different within the HPG system**.**

Besides, through the histogram of volume traded provided, it can be seen that the majority of stock trading occurs at low volume, with the highest frequency on the left side of the chart. This suggests that larger transactions are less common. The apparent decrease in frequency as volume increases may indicate that the market is highly liquid at low volume levels but declines as volume increases. This may also reflect the nature of investors participating in the market, with a large number of retail investors and fewer larger or institutional investors.

**2.Predictive model**

**2.1. Model parameters chosen**

**2.1.1. Arima (0,1,0)**

Via the usage of stepwise search from ‘auto\_ARIMA’, we were able to determine the best parameters to use in this model for HPG Stock Price forecasting, which would be (0,1,0). This specific configuration signifies the following:

* p = 0 (Autoregressive order): The model does not directly incorporate past values of the HPG stock prices to predict future values.
* d = 1 (Differencing order): The model applies first-order differencing to the time series data. This means it works with the changes in price from one period to the next, rather than the raw price values.
* q = 0 (Moving Average order): The model does not include any moving average terms.

Primarily, its simplicity and interpretability make it an attractive choice, as it posits that the best predictor of future Bitcoin price changes is the recent change, with a constant drift factor. The use of first-order differencing effectively addressed the non-stationarity inherent in the original HPG stock price time series. In comparison with other ARIMA models, the (0, 1, 0) model struck a favorable balance between complexity and predictive accuracy. Diagnostic tests further supported this choice: the Augmented Dickey-Fuller (ADF) test confirmed the stationarity of the differenced series, the Ljung-Box Q test did not detect significant autocorrelation in the residuals, and the ARCH test indicated the absence of conditional heteroskedasticity in the model residuals.

**2.2.2. LSTM**

This model consists of 4 main layers: 2 LSTM Layers followed by 2 Dense Layers. Below is a detailed explanation of the functionality of each layer.

* LSTM Layer 1 (128 units, return\_sequences=True): This layer captures and processes sequential information from the input data. The "return\_sequences=True" setting allows this layer to pass its output sequence to the next LSTM layer, and 128 units provide the capacity to learn complex patterns within the sequences.
* LSTM Layer 2 (64 units, return\_sequences=False): This layer further refines the sequential understanding from the previous layer. The statement "return\_sequences=False" means it returns only the final output state of the sequence, effectively condensing the information.
* Dense Layer 1 (25 units): This is a fully connected layer that interprets the features extracted by the LSTM layers. Being a hidden layer, its output is used as input to the next layer, which is our output layer.
* Dense Layer 2 (1 unit) (Output Layer): This final layer takes the refined features from the previous dense layer and produces the final prediction - the HPG stock price.

**2.2. Training and test model results**

**Chart 2.1: Stock Price Prediction with ARIMA and LSTM Models**

|  |  |
| --- | --- |

**Source**: The author, 2024

**Accuracy comparison:**

* **ARIMA(0,1,0):** The predicted values closely follow the actual stock prices. This model successfully captured the trend and the prices in the final part.
* **4-layer LSTM:** The predicted values less closely follow the actual stock prices, capturing the overall trend and fluctuations with relatively better accuracy. While there are still some deviations, especially in predicting the peaks and troughs, the LSTM model appears to provide a more accurate representation of the stock price movement.

**Overall:** Based on the visual comparison of the two charts, the 4-layer LSTM model outperforms the ARIMA(0,1,0) model in terms of accuracy and trend capture. The LSTM model's ability to learn complex patterns and relationships in the data allows it to provide more accurate predictions, especially when dealing with non-linear trends and fluctuations in the stock market.

**2.3. Conclusive remarks and future directions**

Upon comparing the models based on the provided metrics, it is evident that the LSTM model is the superior choice for stock price prediction. The LSTM model's R-squared value of 0.9944 indicates that it explains almost all the variance in the data, demonstrating a significantly better fit compared to the ARIMA model's R-squared value of 0.9399. Additionally, the LSTM model achieves a lower Mean Absolute Percentage Error (MAPE) of 2.1201%, suggesting that its predictions are, on average, more accurate in percentage terms than the ARIMA model's MAPE of 2.6925%. While the ARIMA model presents a slightly lower Root Mean Square Error (RMSE), this is likely due to its tendency to underestimate the upward trend, making RMSE a less critical factor in model selection. Overall, the LSTM model's superior performance in explaining variance and providing accurate predictions underscores its reliability and effectiveness for stock price forecasting.

Further speculations lie on the complexity of each model. The ARIMA(0,1,0) model, with its single active parameter (the differencing order), is inherently limited in its ability to capture complex patterns in the data. It assumes a linear relationship and constant variance over time, which might not hold true in the dynamic and often non-linear stock market environment. On the other hand, LSTM with 4 layers poses as a highly versatile model, proven by its wide application towards other fields, as well as providing specific and accurate prediction on time series analyses as proven by the performance of this model.

In the future, hyperparameter optimization and the incorporation of additional data sources can be considered upon building more complex systems. For the LSTM model, fine-tuning parameters like the number of layers and neurons, as well as experimenting with regularization techniques, can help optimize performance and prevent overfitting. Exploring advanced architectures like bidirectional or stacked LSTMs could further improve its ability to capture complex patterns. For the ARIMA model, incorporating seasonality analysis (SARIMA) and exogenous variables, such as market sentiment or economic indicators, could significantly improve its predictive power. Additionally, both models would benefit from incorporating more historical data and exploring alternative data sources, like news articles and social media sentiment. Implementing these improvements would likely lead to more accurate and reliable stock price forecasts.

**Table 2: Models' Metrics Comparison**

|  | R-squared | MAPE | RMSE |
| --- | --- | --- | --- |
| ARIMA | 0,939856 | 2,692478 | 566,338238 |
| LSTM | 0,99444 | 2,12012% | 657,259575 |